(2)

SECU

AD-A202 964

ENTATION PAGE

Form Approved
OMB No. 0704-0188

1a. RE

1b. RESTRICTIVE MARKINGS

2a. SECURITY CLASSIFICATION AUTHORITY

3. DISTRIBUTION/AVAILABILITY OF REPORT
Approved for public release;
distribution unlimited.

2b. DECLASSIFICATION/DOWNGRADING SCHEDULE

4. PERFORMING ORGANIZATION REPORT NUMBER(S)

5. MONITORING ORGANIZATION REPORT NUMBER(S)
AFOSR-TR- 88-1206

6a. NAME OF PERFORMING ORGANIZATION
Stanford University

6b. OFFICE SYMBOL (If applicable)

7a. NAME OF MONITORING ORGANIZATION
Air Force Office of Scientific Research/NL

6c. ADDRESS (City, State, and ZIP Code)
Stanford, CA 94305

7b. ADDRESS (City, State, and ZIP Code)
Building 410
Bolling AFB, D.C. 20332-6448

8a. NAME OF FUNDING/SPONSORING ORGANIZATION
AFOSR

8b. OFFICE SYMBOL (If applicable)
NL

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
AFOSR-87-0282

8c. ADDRESS (City, State, and ZIP Code)
Building 410
Bolling AFB, D.C. 20332-6448

10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO	WORK UNIT ACCESSION NO
61102F	2313	A4	

11. TITLE (Include Security Classification)
Acquiring Generalizations to Organize Human Databases

12. PERSONAL AUTHOR(S)
Gordon H. Bower & John Clapper

13a. TYPE OF REPORT
Interim

13b. TIME COVERED
FROM 9/1/87 TO 8/31/88

14. DATE OF REPORT (Year, Month, Day)
1988, Oct. 1

15. PAGE COUNT
30

16. SUPPLEMENTARY NOTATION
(None)

17. COSATI CODES

FIELD	GROUP	SUB-GROUP
05	09	

18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
→ key words: personnel management. (JES)

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

Five experiments are briefly described in this report, and plans for three further experiments are set forth. We are investigating the consequences of people forming concepts or categories after they've been exposed to a collection of instances (stimulus objects, patterns, events) for which certain features are highly inter-correlated. One primary consequence is that once such regularities are discovered, they are exploited to greatly simplify the recording of new instances into memory. In particular, new instances come to be recorded simply in terms of their belonging to a familiar category plus having a few distinctive features. We've found strong evidence for this kind of coding of instances. A second consequence is that once the category (correlated features) of an instance is identified, the person can focus his learning efforts on recording the distinctive features of the instances, resulting in better memory for this information. In a short-term memory experiment, we've found strong evidence for this strategy. A third consequence of people learning consistently-correlated features of stimuli is that it affects the way they judge the similarity of two instances. For example, two instances differing on an expected feature were judged more dissimilar than two differing on a variable feature. Furthermore, violations of expected default values are explicitly n d and receive high attentional priority, as evidenced by subjects' reports in our new attribute-listing task.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT
☒ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. ☐ DTIC USERS

21. ABSTRACT SECURITY CLASSIFICATION
Unclassified

22a. NAME OF RESPONSIBLE INDIVIDUAL
Dr. Alfred R. Fregly

22b. TELEPHONE (Include Area Code)
(202) 767-5021

22c. OFFICE SYMBOL
NL

Previous editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE
Unclassified

Objectives and Motivation

In order to act intelligently within its environment, an agent must possess an internal model of that environment (see Craik, 1943; Johnson-Laird, 1983; Gentner & Stevens, 1983). The objective of this research is to investigate how humans learn internal models ("concepts") to characterize general categories of training instances (i.e., objects, events, or situations), and how these models facilitate the acquisition, organization, and retrieval of new information. Understanding how humans acquire and apply models of categories can also provide valuable information about how general conceptual knowledge is represented in long term memory. These issues have long been considered central to a scientific understanding of human intelligence, since concepts are crucial to our abilities to learn, reason, and communicate. Furthermore, the experiments described here may shed light on central functional properties of human learners that could have direct practical applications, by aiding in the design of training programs, instructional materials, computer-assisted learning systems, and the testing and selection of personnel.

Theoretical Background

It is commonly believed that people learn about new objects and events by storing something like a list of features for each such pattern they encounter. But there are at least two limiting factors inherent to this process. The first limitation results from a severe bottleneck in the rate at which people can encode new information into memory (see, e.g., Simon, 1974); this sharply limits how much can be learned in a given interval. The second limitation, common to any intelligent system with a very large memory database, is the difficulty of quickly searching through the database and locating a particular piece of knowledge when it is needed. Human search resources, like encoding resources, are limited; thus, the larger the data base that would be encoded as independent feature lists, the longer this search would take to find a specific bit of knowledge. This fact has received extensive validation; the more independent facts that people are taught about a particular topic or concept, the more time they require to exactly verify any one of them from memory (see J. R. Anderson, 1976, 1983 for reviews of this research). In principle, these two source of competition ought to create major obstacles to our abilities to learn and retrieve information from memory; yet, we do not seem to suffer nearly as much from them in everyday life as laboratory experiments suggest we should. People seem to use efficient strategies to overcome these limitations.

In our Research Proposal, we suggested that humans overcome these limitations by encoding new information in terms of previously-acquired knowledge. We distinguish between two possible memory organizations relating general category knowledge with information about specific instances or exemplars (see Figure 1). In Figure 1a, the learner "tape records" and stores the full listing of features associated with each instance, plus its category membership. No distinction is made between features which are predictable on the basis of knowledge about the general category (henceforth referred to as "category features") and features which are specific to a particular instance (henceforth, "distinctive features"). According to this theory, if the learners have a category model, this will not directly reduce competition for encoding and

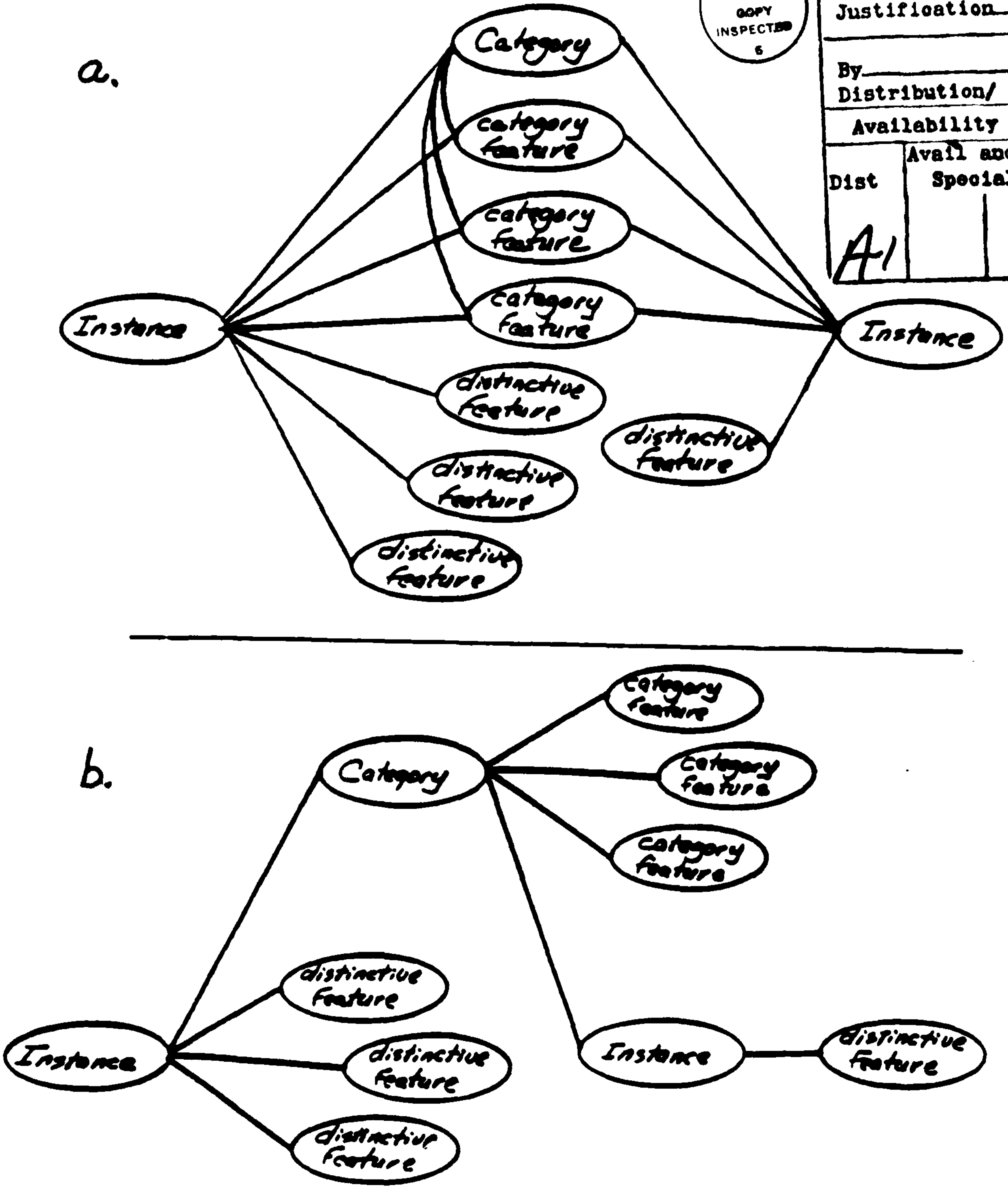

Figure 1. Two possible memory organizations relating general category knowledge to specific information about two instances.

retrieval resources; however, they could use the model at the time of testing to recover category features of the instance which they couldn't retrieve directly. In other words, information stored in the model could be recruited to augment an incomplete feature-list, due either to defective storage or partial forgetting of a particular instance. However, such "trace augmentation" could only be used to improve recall of category features; since distinctive features cannot be inferred from the model, they obviously cannot benefit from this reconstructive process.

Figure 1b depicts a memory organization resulting from a more efficient encoding strategy that we believe people use. In this approach, the learner has re-allocated encoding resources, so that they are not squandered by forming redundant associations from the instance to features which could have been inferred easily from category membership. Consequently, more time would be available for encoding those distinctive features which could not have been predicted generically in advance. Moreover, this "schema-plus-distinguishers" organization facilitates later retrieval of this information: because fewer new associations are learned, there will be less competition for search resources at the time of retrieval. By combining the general knowledge stored in the category model with this reduced representation of instances, all the instance's features can be encoded, stored, and retrieved in the most economical manner possible.

Our hypothesis is that people tend to allocate their attentional resources so as to focus mainly on informative, distinctive features. This hypothesis has several interesting theoretical and empirical implications. In particular, it is incompatible with many current "instance storage" models of category representation (e.g., Medin & Schaffer, 1978; Hintzman, 1986). The major theoretical points, as well as our preliminary experimental research on this topic, were discussed in the Research Proposal. The experiments described below were designed to shed light on these issues, and to develop new empirical methods for their investigation.

Experiment 1: Generalization from Instances

In experiments supported by earlier AFOSR funds (summer 1986), we obtained evidence that people tend to learn instances based on their category membership, as depicted in Figure 1b, rather than "tape recording" a full listing of their features, as in Figure 1a. But the subjects in those earlier experiments were directly taught the generalizations that we wanted them to know for later use; it was therefore important that we also develop an experimental preparation in which we could observe how people learn and apply category models for themselves, using only instances in the absence of explicit category training. Therefore, we conducted an experiment in which subjects were allowed to induce for themselves the shared properties which defined categories of stimuli; we then examined the organization of the memory structures that resulted and how they were used in encoding new instances.

In this experiment, subjects were presented with a series of training instances from two distinct categories. For each category, a set of features consistently appeared in every instance of that category, with different consistent features in the two categories. One category contained three consistent features while the other category had four. Each instance of a given category contained all the consistent features specified by that category, plus one or two additional features that were distinctive to that instance. For purposes of comparison, we also included an equal number of randomly-constructed "control stimuli"; these were matched with the category members in their number of features, but did not possess either of the consistent feature clusters. After repeated experience with instances of both categories, we expected our subjects to learn which groups of features consistently co-occurred across instances. The subjects were not told which category each instance belonged to, or even that there were separate categories in the experiment at all. However, we expected that subjects would discover that many of the stimuli could be characterized as belonging to one of two groups of correlated features; this would be equivalent to their discovering two different categories. Such correlated features should become predicted features (expected defaults) in the subjects' internal models of these categories.

The procedure on each trial will now be described. First, a training instance was presented on a CRT screen. An instance consisted of a series of letters; for instance, a particular instance might contain the letters B, D, Q, and N. This feature list remained on the computer screen for a short period of time (one second per feature), during which the subject was asked to study and try to memorize the instance. After this brief study period, the instance disappeared from the screen and was followed by a short distractor task (a simple arithmetic problem) to reduce short-term memory. Following the distractor task, the subject was asked to free-recall all of the letters in the most recently studied list. Recall was recorded by the subject typing the letters into a computer keyboard.

We expected that subjects would learn the clusters of features which characterized the two categories, and use this knowledge to improve their memory performance. As expected, our subjects showed substantially better memory for the features of category members than for control stimuli of the same length. This result clearly showed that subjects were able to induce category-level knowledge and use it to improve their memory performance. But since both of the memory organizations depicted in Figure 1 can predict some benefit on overall memory performance due to category knowledge, we need to examine more detailed predictions to differentiate them.

The major difference between these theories arise in their predicted effects on memory for an instance's *distinctive* features--those which cannot be directly recovered from a model of the general category. The "tape recorder" theory predicts no benefit of category-learning on memory for distinctive features, relative to the control condition; that theory implies only that the subject can reconstruct the instance's category features from general knowledge at the time of testing. On the other hand, the theory depicted in Figure 1b *does* predict better memory for distinctive features, because the learner is assumed to attend selectively to them during encoding; moreover, these distinctive features experience less competition from other features during the memory retrieval process. In accordance with the latter theory, our subjects showed significantly better recall for distinctive features of category members than for the corresponding features of control stimuli ($p<.01$).

The results of this experiment support our hypothesis that people learn instances in terms of a category model. Their strategy during the study period appears to be this: they quickly identify the category default features in the stimulus; they then neglect these while they selectively attend to and form strong associations with the distinctive features of each instance.

At recall, they recover the category features from the category model, then add the memories of the specific distinctive features of the recent instance they're to recall. These results have opened the way to detailed investigation of the structure of these category models, how they are learned, the strategies by which they are applied, and several other issues of theoretical and practical import.

Experiment 2: The Impact of Knowledge on Perceived Similarity

We claim that people represent instances primarily in terms of a general concept, plus associations to the distinctive features of that specific instance. This distinctive information should receive most attention during the encoding period, whereas the category features should be treated as background. These ideas have several implications regarding people's judgments of how *similar* various category members should appear to each other. By testing and developing these implications, we hoped to use similarity judgments as an analytic tool for illuminating the nature of the underlying pattern representations. In turn, we hoped to uncover new insights about how knowledge affects similarity judgments themselves; this is an important objective because similarity is one of the most fundamental and ubiquitous independent variables in the study of cognition.

Our theory predicts that subjects' judgements of how similar two patterns are will be dominated by the instances' distinctive features--that is, category features will have little impact on subjects' comparisons. This hypothesis leads to the seemingly paradoxical argument that two instances of a given category may often become *less* similar as people become increasingly familiar with that category. A moment's thought reveals that this is actually a commonplace phenomenon associated with expertize: people who become very expert about a particular domain (e.g., expert botanists, wine tasters, dog show judges, etc.) are highly sensitive to differences among the objects in that domain, while taking their well-known commonalities for granted. For instance, a black oak and a red oak are much less alike in the mind of an expert botanist than they are to most non-experts. In fact, our tendency to take for granted what we know about a general domain and to focus attention on the novel aspects of instances is probably what allows us to discover progressively more specific subcategories within that domain, which is one characteristic of experts.

This analysis implies that when an instance of a category violates the subjects' default expectations, such as an instance which has a *missing* category feature, that absence should be highly salient and have a strong effect on similarity. We can use this hypothesis to predict and explain situations in which *decreasing* the number of common features shared by two stimuli leads to *increased* similarity, in direct contradistinction to results ordinarily obtained in similarity experiments (e.g., Gati & Tversky, 1984). This counter-intuitive prediction should be obtained when the subtracted common feature is an expected default for a category to which both instances belong. We hypothesize that deleting a category feature from the stimulus should cause subjects to *add* information to their internal description of the stimulus - namely, that a given expected feature is missing from that stimulus. If this explicit recording of an observed anomaly were to occur for *both* of the stimuli under comparison, the physical deletion operation would have the paradoxical psychological effect of increasing their "common" features. By similar reasoning, deleting an expected feature from one of the stimuli but not from the other would result in the addition of an unusually salient difference (distinctive feature) between them. An expectation failure should produce a very salient feature, so it should cause a greater drop-off in similarity than would a simple difference in the variable features of the two patterns being compared.

The experiment to test these implications consisted of a series of similarity judgements in which college-student subjects rated the similarity of pairs of instances on a 20-point scale. The stimuli were realistic line drawings of fictitious insects ("bugs"), all of which shared a common "base" structure consisting of parts such as head, thorax, abdomen, six legs, eyes, and so on. In addition to this common base, a number of features (e.g., wings, tails, antennae) could be added or removed to construct different instances. Two of these features were consistently presented in all instances (defaults), and two others were presented half the time (variables); additionally, instances were varied along several other attributes to increase the perceived variability of the category. We expected that after several trials, subjects would learn a structural model for the consistently correlated features, treating them like a categoty of stimuli; this model would specify which features were correlated (expected defaults) and which tended to vary across instances. In the midst of this uniform training series, we occasionally presented stimulus pairs in which one or both insects violated the category expectations; such bugs would either be missing an expected default, or they would possess an extra feature not seen in any of the other instances. We were interested in how subjects would rate the similarity of two bugs that were deviant in the same way, in contrast to the way they rated matched, "normal" pairs.

As expected, the results showed that subjects' expectations influenced their similarity judgements. However, violations of defaults had a much larger effect when they served as distinctive features (differences) than when they served as common features. As predicted, we found that pairs in which one member was missing an expected feature (or in which a previously unencountered feature was added) were rated less similar than pairs which differed by a variable feature. In other words, if one insect had wings and the other did not, the effect of this distinctive feature was greater if subjects expected all instances to have wings (or expected none of them to) than if they expected wings to be present or absent equally often. In the data, however, when both test instances were missing an expected feature, their similarity was *not* increased by this shared anomaly -- such pairs were rated the same as pairs in which the defaults were present. This outcome was not as predicted. However, the shared anomaly *did* eliminate the effect of adding common features, an effect that is typically found in similarity experiments. To illustrate, if wings were an expected default, then pairs which had wings were as similar as those which did not; but for subjects who learned wings as a variable feature, pairs which shared this attribute appeared slightly more similar than pairs in which it was missing. Adding an unexpected feature had a slightly larger effect than adding variable features, but this difference did not reach statistical significance.

Although these results clearly showed that subjects' category models were influencing their judgements, we were disappointed that pairs which lacked expected defaults were not rated more similar than normal pairs. Perhaps this was due to subjects weighing distinctive features more heavily in their judgements, than they did common features (seven times as much) -- a typical result with pictorial stimuli (see Gati & Tversky, 1984). In fact, our subjects' reports (and other data) convinced us that most subjects were computing similarity of two bugs by simply counting the differences between the bugs, and largely ignoring common features. This would, of course, wash out the impact of our manipulation of common features. To circumvent this trivial strategy, we designed a second study in which (1) we could independently validate our hypothesis that people note the absence of expected information as a explicit feature of a particular instance, and (2) we could pursue the "common deviation" effect in a situation which minimized those factors that prevented its strong emergence in the previous experiment.

Experiment 3: The Evolution of Conceptual Structure

In this experiment, we pursued our hypothesis that pairs of stimuli which deviate from their category norm in the same way (by lacking the same default feature) should be judged more similar than equivalent pairs of normal stimuli. This result would be important not only for providing further support for our basic attentional and representational hypotheses, but also for showing the importance of shared deviations in guiding people's induction of new categories and subcategories in everyday life. It has been suggested (e.g., by Schank, 1982) that many concepts are primarily defined by clusters of shared exceptions to more general norms. For instance, people have a model or "script" for how to behave in a fast-food restaurant such as MacDonald's (Shank & Abelson, 1977); for many of us, this script is defined mainly in terms of how it violates the standard, sit-down restaurant routine, e.g., by paying before rather than after the meal, by getting food at a stand-up counter. Such intuitions suggest that the increased similarity which results from shared deviance should be an important determinant of our real-world learning.

A second motivation for Experiment 3 was to validate further our claim that people spontaneously notice and encode the absence of expected information. We sought for validation in some behavioral indicators besides similarity judgments. To this end, we developed a new training procedure which allowed us to observe the evolution of a category model as it was being learned. Subjects were shown a series of individual insect stimuli, much like those used in the previous similarity experiment. Once again, there were two features which were consistently present in almost all of the instances (defaults) along with two features which were present or absent equally often (variables). For each instance, subjects were asked to write down a short list of that insect's characteristics. They were told that they need not exhaustively list every feature of the instance, but to list only those which they would need to distinguish it from other members of the same category (e.g., to select it from among several close alternatives on a multiple-choice recognition test). This task had two purposes: (1) to insure that subjects paid full attention to each presented instance; and (2) to allow us to observe whether subjects learned the regularities in the stimuli. Once a category model had been learned, an "ideal subject" in this task should list only the variable features of each stimulus, not its constant (expected) features since only the variable features would be useful for distinguishing difference instances.

In addition to presenting standard category members in this task, we also presented two deviant instances (between the eighth and fifteenth trials). Each deviant insect was missing a different default attribute. We were curious whether subjects would notice this absence, and explicitly list it as a distinguishing descriptor of that instance. Obviously, if subjects spontaneously mention the absence of a particular feature in an instance, one would be forced to conclude that they did, indeed, notice and encode this deviation into their representation of that instance.

Several similarity judgements were interspersed among the attribute listing trials. On two of these trials, the pairs of instances had a common default feature deleted; on another two trials, the insects were identical to those above, except that their default features were left intact. Rather than making a single judgement of overall similarity, as in the previous experiment, subjects compared the two insects according to five different criteria. These criteria were their 1) physical similarity, 2) inferred genetic/evolutionary similarity, 3) inferred behavioral similarity, 4) inferred similarity of habitat, and 5) likelihood of being able to interbreed (another index of inferred genetic relatedness). We suspected that different comparative judgements might be sensitive to different types of information.

The results for the attribute listing task strongly supported our initial hypotheses. As predicted, after the first few trials our subjects mentioned the presence of expected (default) attributes much less frequently than variable attributes. This indicates that they had learned that the presence of the default features could be taken for granted, and that they provided no useful information for differentiating the current instance from other category members. Moreover, this learning occurred very rapidly; after seeing only five instances, the frequency of mention for default attributes had dropped dramatically, from roughly 55 percent on the first instance to 10 percent on the fifth instance (see Figure 2), where it remained thereafter. The overall number of descriptors used dropped by about a constant one-third per trial during this learning interval.

Another result of major interest was that subjects were very likely to notice and report the absence of default features in the two deviant instances. Subjects responded more strongly to the absence of default features (about 75 percent) than they ever did to their presence. The rate of responding for defaults dropped dramatically on the trial following the instance with missing feature, but the rate still remained higher than it had previously been for several trials thereafter. The result suggests that unexpectedly missing feature had temporarily "sensitized" the subjects to noticing and reporting to that feature--much as an unexpected stimulus produces an orienting reflex and temporarily releases habituation from an overly familiar stimulus. This result leaves no doubt that the absence of default features was generally encoded as part of subjects' instance representations.

Interestingly, subjects showed no tendency to rate deviant pairs more similar than normal pairs on any of the five comparison questions, even though we can be quite sure (from the attribute-listing results) that they must have detected the missing attributes. Several alternative explanations of this odd result come to mind. The most plausible is that subjects continued to make their ratings primarily on the basis of the instances' distinctive features, rather than their common features, as in the previous experiment. As described above, several procedural changes were introduced in this experiment to circumvent such a comparison strategy. But the present results, in addition to conversations with our subjects, lead us to believe that these modifications did not achieve their desired result. Apparently, when two well-learned and highly similar visual patterns are presented side-by-side for a similarity judgement, subjects find a "difference counting" strategy very compelling for making their similarity ratings. The next

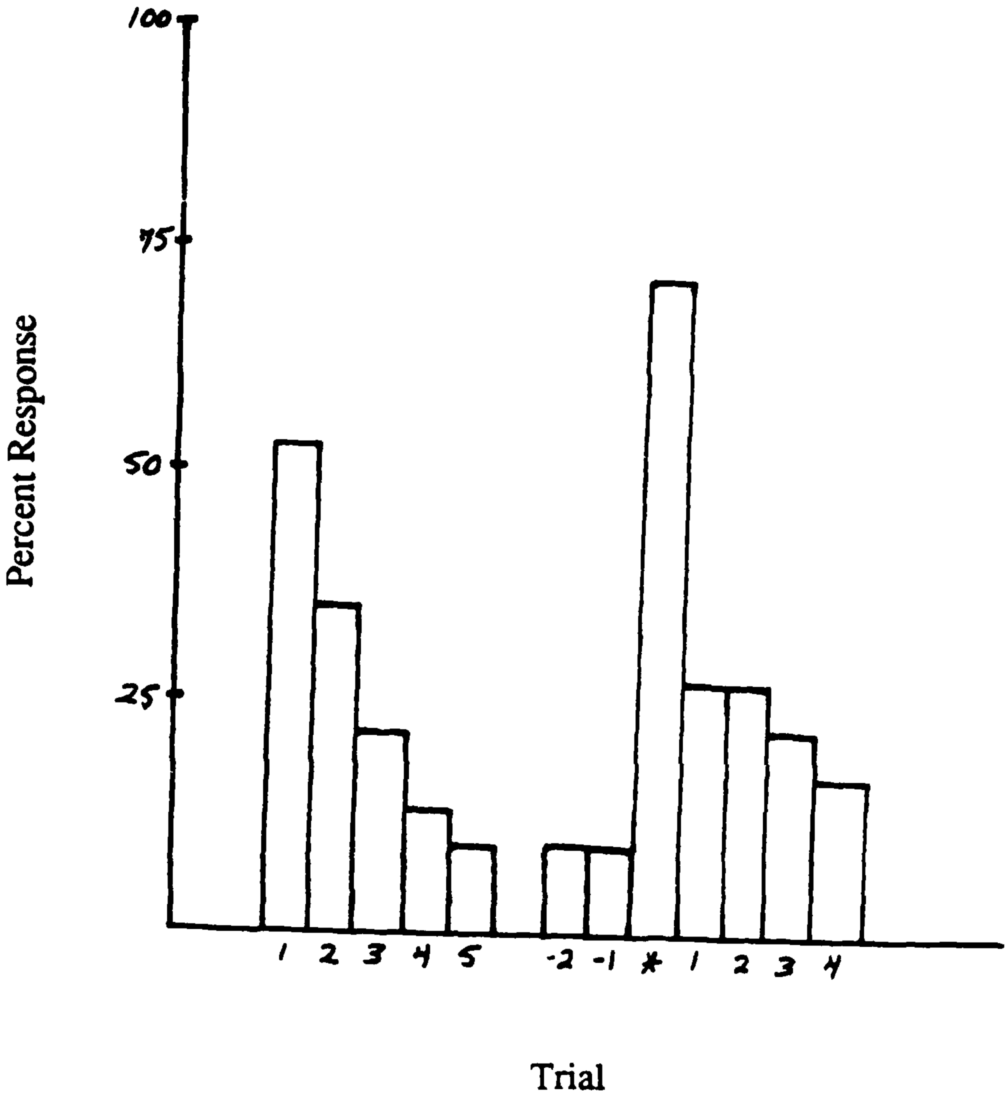

Figure 2. Percent response to default attributes for the first five trials of the attribute listing task in Experiment 3, and for the trials immediately preceding and following the occurrence of an instance missing a default attribute (indicated by an '*').

study was designed to avoid this problem by using a stimulus domain in which people normally weight common features more heavily than distinctive features in their judgments.

Experiment 4 : Shared Deviance and Perceived Similarity

Our first two similarity experiments suggested that a surprising or unexpected feature (e.g., a missing default) has a strong impact as a difference between two stimuli; but those experiments provided no evidence that a missing default had much impact when it is a characteristic shared by the two test instances. However, research by Gati and Tversky (1984) demonstrated with pictorial stimuli that people generally weight distinctive features more heavily than common features in their similarity ratings; conversely, common features had a greater effect on the similarity of verbal stimuli (e.g., descriptions of people, meals, trips, and so on). Thus, our next step in this line of research was to conduct an experiment analogous in most respects to our previous studies, except that the stimuli consisted of *verbal* descriptions rather than *pictures* of objects. We hypothesized that shared violations might increase similarity in such an experiment, in line with our original expectations.

The experimental procedure was similar in most respects to Experiment 2, except that verbal rather than pictorial stimuli were used. The stimuli were lists of disease symptoms supposedly associated with specific medical patients. The subjects rated the similarity of many pairs of such instance-lists, using a 20-point scale as in the previous similarity experiments. The main independent variable in this study was the probability with which different symptoms appeared in the instances. This probability-of-occurrence factor was varied through severel levels; some symptoms appeared very frequently, being listed for over 90 percent of the patients, while others occurred only rarely.

We expected that the weight of a given common feature in the similarity judgement would depend on its probability of occurrence. Features which occurred with a very high probability (i.e., greater than 90 percent) would be learned as defaults in subjects' category models; because of their predictability, such defaults were expected to have little impact on subjects' ratings. By contrast, surprising or exceptional feature values (e.g,. the absence of an expected default, or the presence of a low-probability feature) should be weighted more heavily in subjects' ratings. Thus, pairs in which a default was absent from both instances should have been rated more similar than equivalent pairs in which this feature was present. By similar reasoning, low-probability common features were predicted to increase similarity more than features which occurred with a higher probability. In general, the lower the probability of a given feature value (presence or absence), the greater its predicted impact as a common feature.

The results were somewhat surprising in light of these predictions. Across all experimental conditions, pairs in which a given feature was present were rated more similar than pairs in which it was absent, the usual "common feature" effect found in most similarity experiments (e.g., Gati & Tversky, 1984). We had expected that this typical result would be reversed for defaults, i.e., that the shared absence of a default feature would result in higher similarity than its presence. However, we found exactly the opposite result; eliminating defaults from both pair members caused a large *decrease* in their similarity. In fact, this common-feature effect was significantly *larger* for defaults than for any of the other probability-of-occurrence conditions; by contrast, we had predicted that this effect would be negative, or at least reduced in size (as in Experiment 2). Except for the defaults, the impact of a common feature seemed independent of its probability; for example, features which occurred in 50 percent of the instances had roughly the same weight as those which occurred in under 10 percent. Thus, although our subjects did seem to treat predictable defaults differently from other features which varied unpredictably from trial to trial, the pattern of results was quite different than we had expected.

These results are somewhat counter-intuitive; common sense would seem to imply that two category members which deviate from the norm in the same way ought to appear more similar as a result. In making our original predictions, we assumed that if two instances were members of the same category, then removing a category default from both would increase their perceived similarity. In retrospect, it now seems possible that removing a default from the instances may sometimes eliminate the basis for their common categorization; if this is what occurred, the result would be a reduction in similarity, rather than an increase. This is particularly likely to occur if the categorization of the instances was originally based on a very few consistent features; in this case, removing even one expected feature might change how the instances are categorized. Significantly, there were only two consistently-present features in the present experiment; by contrast, the insect stimuli in the previous similarity experiments bore an obvious structural resemblance to each other. (In general, one difference between pictures and verbal feature lists is that structural overlap is generally much more apparent in the former type of stimuli). Thus, a shift in categorization due to removing a default might have caused the unexpected results obtained in this experiment.

One difficulty which has become apparent in using similarity ratings as a experimental "microscope" for the issues we wish to investigate is that the task seems to impose very little constraint on the strategies which subjects can use to compute their ratings. For example, even when it is apparent that subjects notice certain properties of the stimuli (as demonstrated by other performances, such as the attribute listing task used in Experiment 3), they may not use these properties to compute numerical ratings, or they may use them quite differently in different situations. Thus, we've been forced to conclude that similarity judgments do not seem to provide a very direct index of how people represent the experimental stimuli; they are too strongly influenced by a large number of task-specific factors to be as useful as we had first hoped. By contrast, we have found that memory tasks provide far more constraint on subjects' performance. In our memory experments, the best learning strategy is to give priority to encoding those values which are least predictable on the basis of general knowledge. Since subjects seem to utilize their category models in more predictable and consistent ways in memory tasks than in judgement tasks, the former have so far proven more useful for investigating the variables of interest to us.

Experiment 5: Learning Concepts from Unreliable Data

In most realistic learning situations, the data upon which learners must base their generalizations contain errors and exceptions. In this experiment, we wished to study peoples' ability to learn relatively idealized concept models from "noisy" data, where generalizations were somewhat unreliable because usually consistent features were sometimes replaced or missing. This experiment should establish the generality of our earlier results and boundary conditions for their application; in addition, its results should help us discriminate among the various hypothetical processes by which concept models might be learned and applied.

A recall procedure similar to that in Experiment 1 (Generalization from Instances) was used. The stimuli were sequences of six letters; each position in the sequence can be thought of as an attribute, and the letters which fill that position as the alternative values of the attribute. All attributes were trinary, i.e., there were three possible letters which could occur in a given position. On each trial, the subject was presented with a single instance (six-letter string) for a brief study period, followed by a 15-second distractor task to reduce short-term memory, and then a recall test. Each subject was exposed to instances of a single concept, characterized by both "Consistent" and "Variable" attributes. For the Consistent attributes, one value occurred more frequently than the other two; this was the modal (default) value of that feature. In contrast, all three values of the Variable attributes occurred equally often. The major independent variables were (1) the *reliability* (probability of occurrence) of the modal values of Consistent features (60, 70, 80, or 90 percent); and (2) the *ratio* of Consistent to Variable features characterizing the concept (four Consistent and two Variable, versus two Consistent and four Variable). A control condition was also included in which all six attributes were Variable. Thus, a total of nine conditions were to be run in a between-subjects experiment design.

We were interested in how these factors affect peoples' ability to learn and apply idealized concept models in the face of noisy data. Given such a model, remembering any instance need only involve (1) encoding the values of its Variable features, and (2) encoding any non-default values of its Consistent features. This implies that subjects should best remember the Variable features of prototypical instances, whose Consistent features all have their default values. Each exceptional value of the Consistent features competes with learning and remembering the Variable features; hence, the more such exceptions are present, the poorer memory should be for the Variable features. This memory benefit for prototypical instances should index the extent to which subjects rely on a category model to learn the instances. By examining how this benefit varies with the reliability and number of default attributes, we may be able to characterize how these factors affect the formation and utilization of concept models.

At the time of this writing, a partial version of this experiment has just been completed at Brooks Air Force Base, and we are awaiting the data. Provided that this data "checks out" (i.e., subjects understand the task, are performing at a reasonable level of accuracy, and so on), we will then proceed with the complete experiment. We anticipate completion of this experiment within the next six to eight weeks.

Forecast

We are currently planning a number of experiments, some of which are extensions of the research reviewed above or originally described in our Research Proposal, and some of which are based on ideas developed over the last year. We anticipate completing the following three within approximately the next six months. We hope to conduct one or two of these experiments at the LAMP laboratory at Brooks Air Force Base in San Antonio.

Proposed Experiment 1: The Organization of Natural Knowledge Bases

The purpose of this proposed experiment is to develop a general procedure or set of diagnostic indices for determining the structure of naturally existing databases (so-called "semantic memory"). This is an important objective for validating the everyday relevance of conclusions based on laboratory-produced knowledge. Equally important, it could help us to discover important new questions that might be missed if we only studied artificial knowledge structures. Our main objective in this research is to discover what types of facts tend to be stored together (directly associated with each other) in semantic memory. Previous attempts to accomplish this goal (e.g., Collins & Quillian, 1969) were flawed in several respects, but we believe that we have developed a procedure which avoids these difficulties and will allow us to make solid conclusions about long-term memory organization.

The general procedure is as follows: On each trial, the subjects will be presented with one or more facts, all supposedly pertaining to a particular instance of a familiar category (e.g., a bird, chair, restaurant visit, etc.). This feature list will include: (1) the category label for that instance; (2) zero, one, or two facts predictable from this category membership, and (3) zero, one, or two facts *not* predictable from category membership (or presumed not to be directly associated with it in semantic memory). For example, the subject might be told the instance under consideration is a *bird*, has *feathers*, has a *beak*, is *yellow*, and is in the *park*. The last two properties are irrelevant (or not defining), and the middle two are relevant (defining) to the *bird* category. Following this study period, a test statement will be presented; the subject must decide from memory as rapidly as possible whether this fact was in the list they had just studied. For example, the subject might be asked whether it was stated that the instance was *yellow*. Both the speed and accuracy of subjects' decisions will be recorded.

Our main hypotheses concern the effects of category-relevant versus category-irrelevant facts on the speed and accuracy with which subjects can retrieve the instance's category label. Our predictions can be understood by referring to Figure 3, which depicts an instance associated in memory with several presented facts. By this depiction, associations with irrelevant facts should compete for search resources (activation) with the association from the instance to the category label. Because such irrelevant associations reduce the amount of activation which reaches the category label, they should reduce its availability from memory. The effects of category-relevant facts should be somewhat more complex. Due to their prior associations in semantic memory with the category label, activation of these facts should also activate the category label; but this activation arrives indirectly and consequently should be somewhat dissipated relative to when the category label is presented alone. Therefore, adding only

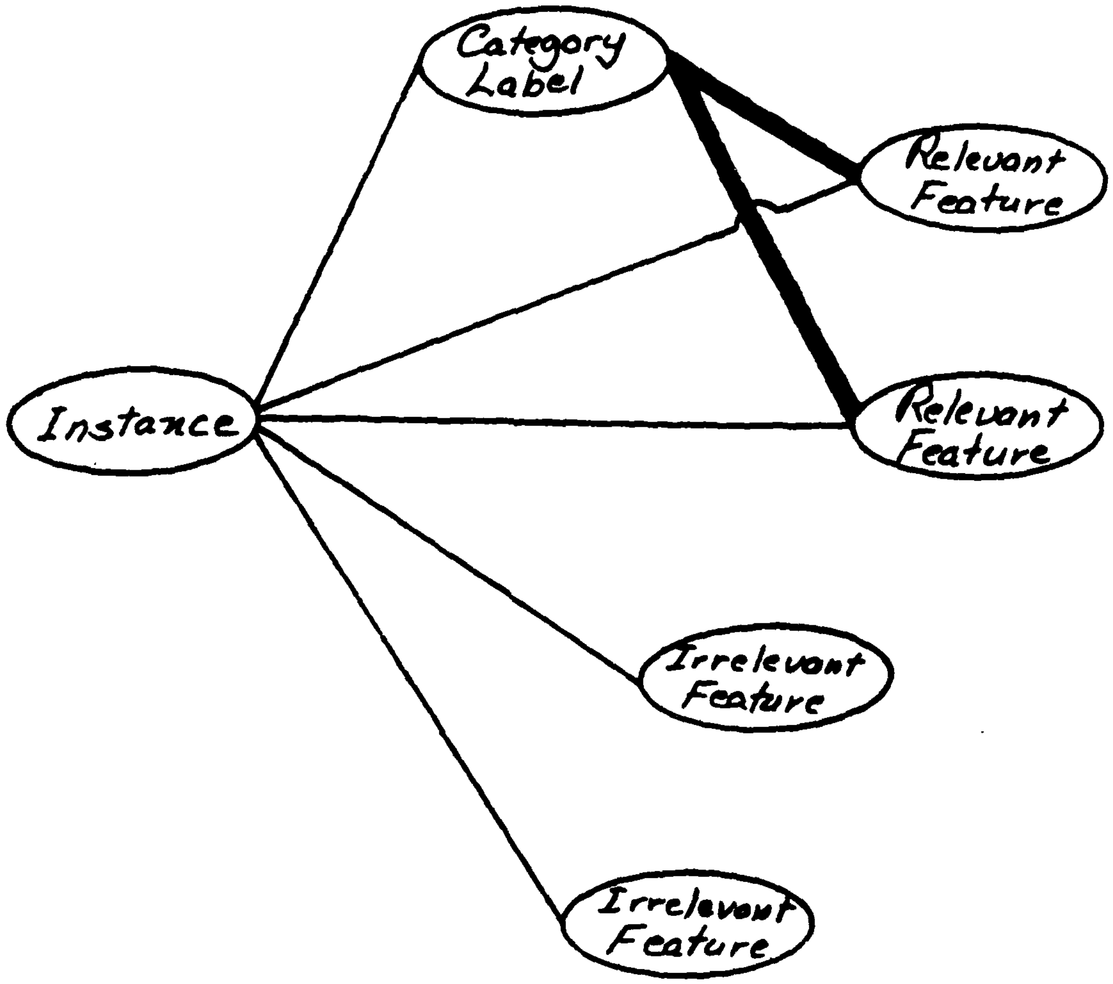

Figure 3. Associative structure representing an instance plus its associated feature list. Previously existing category-to-feature associations are indicated by boldface links.

relevant facts ought to slightly reduce the availability of the category label, although much less than adding irrelevant facts would. On the other hand, if irrelevant facts have also been presented, then adding relevant facts should tend to siphon activation away from these other facts and back to the category label, reducing the severity of the competition from the irrelevant facts. This implies that adding relevant facts should increase the availability of the category label *if* irrelevant facts have also been presented; but without such irrelevant facts, relevant facts should either have no measurable effect or slightly impair category retrieval.

These predictions assume that the relevant facts have prior associations in long-term memory with the category, and that the irrelevant facts do not. The experiment will test this representational assumption; the predicted results should obtain only if it holds for the categories we investigate. Examining such patterns of facilitation and interference will provide a general procedure for investigating the organization of semantic memory. Irrelevant facts will be of several types, including: (1) specific facts about an individual, not predictable from any existing categorization (our first planned experiment of this type); (2) facts predictable from a superordinate level of categorization, but which in theory should not be directly associated with the target category (e.g., "has lungs" or "has bones" for a bird), and (3) facts predictable from a subordinate level of categorization (e.g., "is nocturnal" and "is wise" for owls as a type of bird). Such experiments should demonstrate that information pertaining to superordinate or subordinate categories tends to be stored at its appropriate level of categorization and is not directly associated with the target category, as hierarchical models of semantic memory organization suggest. We could also use this technique to demonstrate that our knowledge of specific domains is *not* organized in this fashion; this would allow us to develop theories which specify the particular conditions under which different types of memory organizations tend to evolve. If successful, this technique could become a valuable new tool for the investigation of semantic memory.

Proposed Experiment 2 : Abstracting Regularities from Training Instances

In our Research Proposal, we argued that instance-based learning models do not provide straightforward capabilities for the kind of knowledge organizations that we have created in our experiments. To arrange another test, we note that different theories assume that learners use different aspects of the training instances to guide their performances. Instance-based models assume that classification performance is based on computing the similarity of the new instance *independently* to each instance stored in memory, or to some remembered subset of these instances (e.g., Estes, 1986; the Minerva model of Hintzman, 1986; the Context model of Medin & Schaffer, 1978). This *Independent Instances Assumption* implies that learners should not be sensitive to how much the instances in memory overlap with each other, but only to how similar each memory-instance is to the new instance to be classified. To illustrate, we will denote experimental stimuli by vectors of binary values. Each element in these vectors stands for a particular feature of a given stimulus. Taking a stimulus denoted by 111111 as a base, the stimuli 111001, 111100, and 111010 all differ from 111111 by two feature values. Note that stimuli 011011, 101101, and 110110 also differ from 111111 by two features. But the former three instances all overlap with 111111 on the *same* three features, whereas the latter three overlap on different features. Our theory expects that a schema would be formed for the former set of stimuli, which have a *consistent* set of overlapping features, but that no schema would be formed for the latter set. However, a strict instance-based model would not predict any difference in how easily 111111 would be categorized following subjects' exposure to the two sets of training stimuli. While standard Instance-Storage models are only sensitive to instance similarity, models that compute generalizations directly capture and make use of the consistently high feature-correlations in a given domain (Elio & Anderson, 1981).

The most straightforward way to evaluate the Independent Instance assumption is to test whether people are sensitive to the indicated type of overlap among instances in memory. Instance-Storage models attempt to account for transfer performance (e.g., on a following classification task) entirely on the basis of summating the similarity of the transfer instance across individual instances in memory. However, it is possible to hold this factor (similarity) constant while separately manipulating the consistency of the overlap among the training instances. If subjects' performance in remembering or classifying a new instance differs depending on this overlap manipulation, then we can conclude that they are learning and using information other than that test instance's similarity to previously stored instances.

We propose an experiment that directly tests the Independent Instance assumption. Subjects will first be taught about two categories by showing them three or four exemplars of each. The instances of one of these categories will overlap on a consistent subset of their features (e.g., 111010, 111001, 111100), while the instances of the other category will lack such a consistent feature group (e.g., 011011, 101101, 110110). We will refer to the former category as the *Consistent* category, while the latter will be designated the *Inconsistent* category.

Following exposure to these instances, a transfer task will be administered to assess what knowledge subjects have extracted from the training stimuli. Subjects will be presented with a single test stimulus, and will choose which of the two training categories this stimulus best exemplifies. A single pair of three- or four-instance training categories plus a transfer test can be thought of as comprising a single trial. Each subject will receive many such trials, with different stimuli each trial. By manipulating the nature of the categories and their relation to the test instances, we can diagnose what subjects learn about each category, and how this serves as a basis for their categorizations.

On some trials, the test stimulus will have the same average similarity to both categories. For example, the stimulus 111111 differs by two features from each stimulus in the two training sets above. If subjects learn the default features of the Consistent category, then they should categorize the test stimuli primarily on the basis of this feature group. If they do not discover the consistently correlated elements, then they should choose Inconsistent and Consistent categories equally often (assuming the two training sets are equally similar to the test stimulus). Moreover, if subjects base their classification performance mainly on the consistent feature groups, then a training set containing such consistencies should be preferred over the alternative even if the latter has *greater* average similarity to the test stimulus. We will compare subjects' classification preferences when the training categories are equally similar to the test stimulus, and also when the Inconsistent category is more similar to the test stimulus than the Consistent category. We predict that subjects will prefer to classify on the basis of featureal consistency, and will only use similarity when they have no other basis for making their choices.

This paradigm is very analytic, and several fine-grained issues could be investigated using it. Such investigations would involve varying the nature of the training categories (the type of stimuli they contain and the manner in which they are presented), and observing how these manipulations affect subjects' categorization preferences. For example, when subjects are forced to classify on the basis of similarity (i.e., when neither category has a consistent feature group), will they classify mainly on the basis of average similarity to all stored exemplars, or on the basis of high similarity to a single memory-instance? How will this strategy be affected by the number of instances they have stored in memory for each category, and by how similar these instances are to each other? Such issues can be easily investigated within the proposed paradigm.

Proposed Experiment 3 : Spontaneous Induction of Default Hierarchies

Much of our knowledge of the world is stored in complex *default hierarchies*, i.e., organized around expectations derived from subordinate and superordinate relations among different concepts (see, e.g., Holland, Holyoak, Nisbett, & Thagard, 1986). For example, categorizing an object as a robin makes available a rich set of default expectations based on that specific categorization, as well as more general expectations derived from the knowledge that robins are birds, that birds are a type of animal, and so on. Since people seem to construct organized default hierarchies when learning about most real-world domains, rather than merely acquiring collections of unrelated single concepts, it is important to understand the processes by which such learning occurs and the kind of knowledge base that results. Investigations in this area will contribute to a general theory of inductive learning. Additionally, they should be applicable to predicting human performance in applied domains, by elucidating the factors which control inductive performances.

The attribute listing task (introduced in Experiment 3) provides an excellent procedure for observing the evolution of default expectations on a trial-by-trial basis. In this proposed experiment, we plan to extend and develop this promising task to study the spontaneous induction of relatively complex default hierarchies. The stimuli for this experiment will consist of pictures of fictitious insects, similar to those used in Experiments 2 and 3. In those experiments, the stimuli were all instances of a single general concept. By contrast, stimuli in the present experiment will be partitioned into two general (superordinate) categories, each of which can further be divided into two more specific (subordinate) categories. An abstract depiction of such a stimulus set is shown in Figure 4. In this illustration, the categories are represented by vectors of binary values. The elements in these vectors correspond to specific stimulus attributes, such as wings, eye color, and so on. Each "0" or "1" bit indicates which value (e.g., large/small wings) of the corresponding attribute is the default value for that category. An "X" value indicates that the attribute varies randomly across different category members. The superordinate categories are distinguished by different default values on four consistent attributes. To illustrate, if these four attributes were assumed to correspond to wings, tails, abdominal markings, and eye color (in that order), then category A insects might all have long tails, small wings, stripes, and white eyes, while category B insects would have large wings, short tails, spots, and black eyes as their default properties. The subordinate categories are

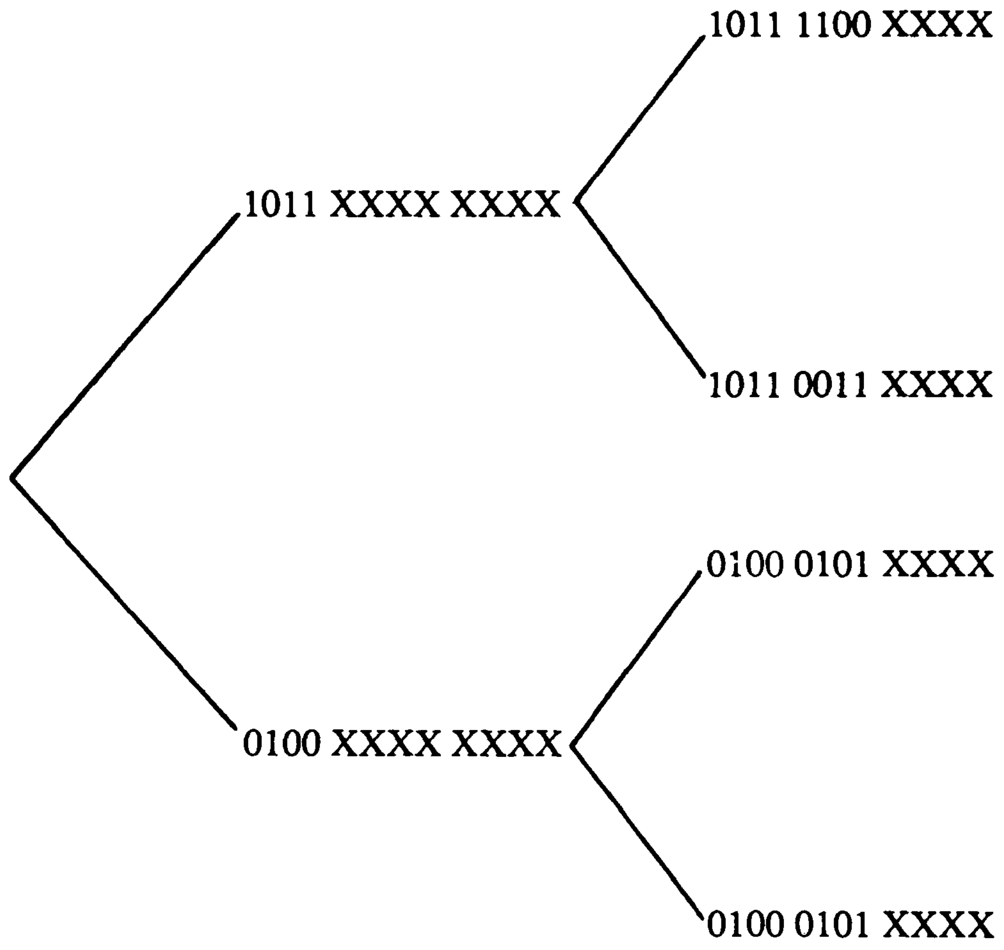

Figure 4. An abstract depiction of the stimulus set to be used in Proposed Experiment 3.

distinguished in a similar manner, by different default values on another set of four attributes. These subordinate categories inherit the default properties of their superordinates; thus, category A1 and category A2 would hav: the same default values on four of their attributes, and contrasting values on four others.

The procedure will be similar to the attribute listing task introduced in Experiment 3. Subjects will be shown a series of individual stimuli, each a member of one of the categories described above. For each presented instance, the subjects will be instructed to write down whatever features of that insect they feel they would need to reconstruct it for a later recall test (e.g., to draw the instance from memory with the aid of this feature-list). Within this memory constraint, they will be urged to minimize the number of characteristics that they list for each insect, i.e., to imagine that each feature they list costs them a dollar. As in previous experiments, subjects will not be told which category each instance belongs to, or that the stimuli could be divided into separate categories at all. However, we expect that subjects will discover this for themselves, and that this knowledge will be reflected in the features which they list for each instance.

The attributes listed on each trial will serve as an index of what subjects have learned about the consistencies in the stimulus set up to that point in time. Eventually, most subjects should list only those features which are not predictable defaults. We expect that the subjects will first discover which attributes differentiate the two superordinate categories. As a result, they should begin to drop these predictable attributes from their listings, continuing to mention only one to indicate the category membership of each instance. For example, rather than explicitly listing all four superordinate defaults, a subject might simply characterize each instance as "large winged" or "small winged". Thus, we expect that a single attribute will come to serve as a proxy in subjects' listings for all the default features of a given category.

After they have learned the two superordinate concepts, the subjects should next discover further regularities which give rise to the four subordinate concepts. This learning should be indicated by a drop-off in the frequency with which subordinate defaults are included in subjects' listings. Thus, we again predict that a single value will serve as a proxy for a larger group of correlated subordinate defaults in subjects' listings. After their learning is complete, the number of properties which subjects should, in principle, need to list for a given instance drops from an initial high of twelve to a final low of only six. This represents a considerable savings in the amount of information subjects would need to record about a given instance in order to fully reconstruct it at a later time.

We think the attribute listing task will be sensitive enough to reveal the evolution of such conceptual hierarchies. If so, it will be one of the first clear revelations of how such hierarchies are learned and utilized in organizing large data-bases. And that is one of the aims of this project.

References

Anderson, J. R. (1976). *Language, memory, and thought.* Hillsdale, N. J.: Lawrence Erlbaum Associates.

Anderson, J. R. (1983). A Spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior*, *22*, 261-295.

Collins, A., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *8*, 240-247.

Craik, K. (1943). *The nature of explanation*. Cambridge: Cambridge University Press.

Elio, R., & Anderson, J. R. (1981). The effects of category generalization and instance similarity on schema abstraction. *Journal of Experimental Psychology: Human Learning and Memory*, *7*, 397-417.

Gati, I., & Tversky, A. (1984). Weighting common and distinctive features in perceptual and conceptual judgements. *Cognitive Psychology*, *16*, 341-370.

Gentner, D., & Stevens, A. L. (1983). *Mental models.*. Hillsdale, N.J.: Erlbaum.

Hintzman, D. L. (1986). Schema abstraction in a multiple-trace memory model. *Psychological Review*, *93*, in press.

Holland, Holyoak, Nisbitt, & Thagard. (1986). *Induction: Processes of inference, learning, and discovery*. Cambridge, Mass.: MIT Press.

Johnson-Laird, P. N. (1983). *Mental models: towards a cognitive science of language, inference, and consciousness*. Cambridge: Harvard University.

Medin, D. L., & Schaffer, M. M. (1978). A context theory of classification learning. *Psychological Review*, *85*, 207-238.

Schank, R. C. (1982). *Dynamic memory*. Cambridge: Cambridge University Press.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals, and understanding*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Simon, H. A. (1974). How big is a chunk? *Science*, *183*, 482-488.